Application test of PCA based improved K-means clustering algorithm in analyzing NGO assistance needs in less developed countries

Linhui Qin

Department of Accounting and Finance, Middlesex University, London, The Burroughs, Hendon, London, NW4 4BT

qlhkv98@163.com

Abstract. In today's society where the amount of data is increasing by Peta Byte (PB) or Exa Byte (EB), it is an era of big data explosion, but there are also some unlabeled data or unstructured data. Compared with complex supervised learning, unmarked unsupervised learning has great potential and value in social development. The clustering algorithm K-means is one of the commonly used algorithms in unsupervised learning. However, after studying the shortcomings of K-means itself, a problem is found that the dimension attribute of the data set must be converted into a numeric type by means of arithmetic average to measure the distance. Different random selection will have a certain degree of influence on the final clustering results, and eventually lead to the decision deviation is too large. Especially for high noise points, multidimensional, nonlinear social big data. In order to solve this problem, the theme of this paper is the application test of PCA based improved K-means clustering algorithm in analyzing NGO assistance needs in less developed countries. First, read and clean up the national data of 167 less developed countries. Secondly, data visualization and data preparation are carried out to re-scale. The principal component analysis algorithm is used to analyze and deal with outliers. Clustering trends are analyzed by combining a k-means model determined by scores obtained from the Hopkins statistical test with a list of countries ultimately in need of assistance. Finally, it can be tested that PCA data cleaning can effectively reduce data noise and improve the clustering effect.

Keywords: PCA, K-means algorithm, data mining, visualization, Hopkins statistics.

1. Introduction

In 1959, American artificial intelligence pioneer Arthur Samuel proposed the term "machine learning". In the following decades, benefiting from digitized information, the field of machine learning has witnessed a burst of development around statistics and probability [1]. Up to now, with the arrival of the era of big data, intelligent analysis has become an important means of information data analysis for countries, non-governmental organizations, enterprises, people and so on in different dimensions of global development decision-making. There are many forms and standards of machine learning. It is mainly divided into three categories, namely supervised learning, unsupervised learning and reinforcement learning [2]. K-Means clustering is one of the most common clustering algorithms in the field of machine learning. However, due to the shortcomings of K-means itself, the dimensional attributes of the data set must be converted into numerical types through arithmetic means to measure the distance. Different random selections will have a certain degree of influence on the final clustering results, and eventually lead to excessive decision-making bias [2][3]. Especially high noise points, multidimensional and nonlinear social big data have many error values that affect clustering [4]. From the perspective of data cleaning, this paper focuses on the application of improved K-means clustering algorithm based on PCA. Combined with Hopkins statistics, the national information data of 167 less developed countries in the world. Perform data reading and cleaning. The cleaning method of this study is to minimize the data loss by maximizing the variance of PCA [5]. Dimension reduction reduces the impact of data noise and facilitates data visualization and rescaling. The reliability of data clustering was determined by combining the test scores obtained from Hopkins statistics. Improve the scene applicability of K-means clustering algorithm. In fact, unsupervised learning technology is more important for data science research in the future of social development. The world's less developed countries have different national conditions. There are often non-objective factors, which will inevitably affect the validity of the data itself.

2. Improved K-means clustering algorithm based on PCA combined with Hopkins statistic test

2.1. Methodology

2.1.1. PCA. PCA is a linear dimension reduction algorithm, which is usually used for data preprocessing. Its goal is to use variance to measure the difference of data, and project the highly different high-dimensional data into the low-dimensional space for reduction and noise reduction [6]. The principle is as follows:

First of all, there are m sample points in the R^dspace. Expressed as X= $\{x_1, x_2, x_3... x_m\}$. These points represent the characteristics of the data of less developed countries, such as GDP. Each data x_i has d dimension and x_i belongs to R^d.

$$x_i = \begin{Bmatrix} x_i^1 \\ x_i^2 \\ \vdots \\ x_i^d \end{Bmatrix}_{d*1} \tag{1}$$

The formula (1) can be obtained

$$X = \{x_1, x_2, x_3 \dots x_m\}_{d*m}^T = \begin{bmatrix} x_1^1 & \cdots & x_1^d \\ \vdots & \ddots & \vdots \\ x_m^1 & \cdots & x_m^d \end{bmatrix}_{m*d} \tag{2}$$

Where x_i^j represents the feature of the j dimension of the i sample, and separately listed vector x_i is the d dimension. Each row represents a sample. The x_i data is normalized to ensure that its mean is 0

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i = \begin{Bmatrix} \bar{x}^1 \\ \bar{x}^2 \\ \vdots \\ \bar{x}^d \end{Bmatrix}_{d*1} \tag{3}$$

Calculate the mean value of corresponding elements for all data sample point column vectors x_i to obtain the mean value column vector $\bar{x}$ of d dimension. Then subtract the mean value from the original sample data to obtain the normalized result.

$$\mathrm{x_i} = \mathrm{x_i} - \bar{\mathrm{x}} = \left\{\begin{matrix} \mathrm{x_i^1} - \bar{\mathrm{x}}^1 \\ \mathrm{x_i^2} - \bar{\mathrm{x}}^2 \\ \vdots \\ \mathrm{x_i^d} - \bar{\mathrm{x}}^\mathrm{d} \end{matrix}\right\}_{\mathrm{d*1}} \tag{4}$$

Dimension reduction compression was carried out for data sample $\mathrm{R^d}$ belonging to $\mathrm{x_i}$.

The result is that $\mathrm{R^{d'}}$ belongs to $\mathrm{z_i}$, (d'< d). The process of obtaining $\mathrm{z_i}$ from $\mathrm{x_i}$ can be expressed as follows: linear transformation of matrix underweight $\mathrm{W} = \{\mathrm{w_1}, \mathrm{W_2} \ldots \mathrm{w_{d'}}\}$. This is the result of our dimensionality reduction, and this is the goal of PCA.

$$\mathrm{z_i^j} = \mathrm{w_I^T} = \left\{\begin{matrix} \mathrm{w_j^1} \\ \mathrm{w_j^2} \\ \vdots \\ \mathrm{w_j^d} \end{matrix}\right\}_{\mathrm{d*1}}^{\mathrm{T}} \cdot \left\{\begin{matrix} \mathrm{x^1} \\ \mathrm{x^2} \\ \vdots \\ \mathrm{x^d} \end{matrix}\right\}_{\mathrm{d*1}} \tag{6}$$

Where, $\mathrm{z_i^j}$ is the element value of the J-dimension corresponding to the original sample data $\mathrm{x_i}$ after dimension reduction; And $\mathrm{w_j}$ is a d-dimensional column vector.

Starting from the maximum projection variance, and when $\left|\left|\mathrm{w_j}\right|\right|_2^2 = 1, \mathrm{w_j^T} * \mathrm{x_i}$ is the projection of sample data points in the direction of this matrix. Because the projection from α to β is |α|cos θ, and the dot product of the vector is α * β equals |α| |β| cos θ, and when |β|equals 1, the dot product is the projection.

Our goal is to find the right W to maximize the variance of this projected $\mathrm{w_j^T} * \mathrm{x_i}$, that is, to divide the sample points as far apart as possible on the new basis. And then we have the sample variance, and notice that the mean has normalized to 0.

$$\mathrm{S} = \frac{1}{\mathrm{m}-1} \sum_{\mathrm{i}=1}^{\mathrm{m}} \left(\mathrm{w_j^T x_i} - \mathrm{w_j^T}\bar{\mathrm{x}}\right)^2 \tag{7}$$

By calculation, we can get $\mathrm{w_j^T S w_j}$ the variance of the projection turns out to be the covariance of the sample matrix.

Then the optimization target is obtained through the sum calculation.

$\arg\max \mathrm{w_j^T S w_j}$

$$\mathrm{s.t.}\ \mathrm{w_j} * \mathrm{w_j^T} = 1 \tag{8}$$

Next, use Lagrange day multiplication

$$\mathrm{L}(\mathrm{w_j}, \lambda) = \mathrm{w_j^T S w_j} + \lambda(\mathrm{w_j} * \mathrm{w_j^T} - 1) \tag{9}$$

And then for each of these $\mathrm{w_j}$, Find the partial derivative $\frac{\partial \mathrm{L}}{\partial \mathrm{w_j}} = 0$

$$\mathrm{S w_j} - \lambda \mathrm{w_j} = 0$$

$$\mathrm{S w_j} = \lambda \mathrm{w_j} \tag{10}$$

In formula (10), $\mathrm{w_j}$ is the eigenvector and λ is the eigenvalue.

Back to the optimization goal again

$$\mathrm{w_j^T S w_j} = \mathrm{w_j^T}\, \lambda \mathrm{w_j} = \lambda \tag{11}$$

This means that the result of maximizing the variance is the eigenvalue of the covariance matrix of the data sample. You can understand that the eigenvalues in some sense represent the meaning of variance. And so on, every $\mathrm{w_j}$ that you get after that has to be orthogonal to each other dot by 0.

Therefore, it can be generally understood as finding the covariance matrix S of data sample X, which is a real symmetric matrix, so it must be similar to diagonalization. Find its largest eigenvalue λ, which is the largest projection variance, and its corresponding eigenvector w_j, which is the linear multiplication factor. And finally, choose the largest eigenvalue. Generally speaking, the solution is directly solved by singular value decomposition method at present, because the process of singular value decomposition of data matrix X will also construct XX^T. That's the covariance matrix.

At this point, we only need to consider one of the parameters in the multidimensional data analysis, so that a feature (dimension) can be reduced. The PCA method is designed to accomplish this process using a rigorous data method.

Algorithm steps of PCA.

There are m pieces of N-dimensional data.

In the first step, the original data is formed into an n row m column matrix X by columns.

In the second step, zero means each row of X (representing a property field), that is, subtract the mean of the row.

Third, find the covariance matrix $C = \frac{1}{m} * XX^T$

The fourth step is to find the eigenvalues and corresponding eigenvectors of the covariance matrix.

In the fifth step, the eigenvectors are arranged into a matrix from top to bottom according to the corresponding eigenvalues, and the first k rows are taken to form the matrix P.

In the sixth step, Y=PX is the data after dimension reduction to k dimension.

2.1.2. Hopkins statistics. Hopkins statistics, spatial statistics, test for spatial randomness in a given data set D, which can be viewed as a sample of a random variable *o*, and we want to determine to what extent *o* differs from a uniform distribution in the data space [7]. The principle is as follows.

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} x_i} \tag{12}$$

y_i is calculated by generating another uniform distribution, sampling data from this distribution, and calculating the closest distance to the data point in dataset D to get y_i. Therefore, if D is also uniformly distributed, then $\sum x_i$ and$\sum y_i$ will be close, and H will be close to 0.5. But the data in D is not uniformly distributed, so $\sum y_i$ will be greater than $\sum x_i$,H will be greater than 0.5, and there may be clusters.

In this study, D is the national conditions data set of less developed countries. This method is used as a statistical hypothesis test for clustering trends, where the null hypothesis is that the data is generated by the Poisson point process and therefore distributed uniformly and randomly. A value close to 1 tends to indicate that the data is highly clustered, with random data often having a value around 0.5, while evenly distributed data tends to have a value close to 0.

2.1.3. K-means. The input parameter of the k-means algorithm is based on the number of generated clusters k.

In the first step, then divide n objects into k groups. n), each group represents a cluster.

In the second step, k objects are randomly selected as the initial clustering center

In the third step, the remaining objects are assigned to the most similar cluster according to their similarity (distance) with these clustering centers. Then the clustering center of each new cluster is calculated.

Fourth, generally the mean square error is used as the judgment rule in the process of beginning convergence. [8].

The specific definition is as follows.

$$E = \sum_{i=1}^{k} \sum_{p \in c_i} |p - m_i|^2 \tag{13}$$

In the formula, E is the sum of the mean square error of all data sets after cleaning. P is a point in the space of the object; m_i is the mean value of cluster c_i.

3. Application to the analysis of NGO assistance needs in less developed countries

Applying the overall process starts with effective data cleansing and visualization by reading and understanding the data. Principal component analysis was performed when the data was fully prepared. Finally, the model was established by Hopkins statistical test to complete the final analysis and get the results.

3.1. Reading and understanding the data

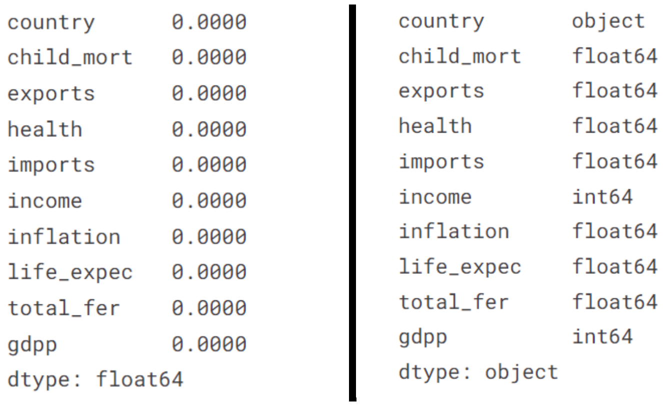
This study focuses on poverty eradication through international aid NGOs, providing basic living facilities and relief to people in backward countries when disasters and natural disasters occur. To do so, data analysis was required for 167 less developed regions by focusing on relevant issues in the countries most in need of assistance. According to Table 1, a total of 9 characteristic quantities are provided in the collected data set for research. The results of the analysis are then provided to NGO decision makers.

Table 1. Characteristics of the data examined.

	Column Name	Description
0	country	Name of the country
1	child_mort	Death of children under 5 years of age per 1000 live births
2	exports	Exports of goods and services. Given as % age of the Total GDP
3	health	Total health spending as % age of Total GDP
4	imports	Imports of goods and services. Given as % age of the Total GDP
5	Income	Net income per person
6	Inflation	The measurement of the annual growth rate of the Total GDP
7	life_expec	The average number of years a new born child would live if the current mortality Patterns are to remain the same
8	total_fer	The number of children that would be born to each woman if the current age-fertility rates remain the same
9	gdpp	The GDP per capita. Calculated as the Total GDP divided by the total population

3.2. Data cleansing

According to Figure 1. After the data set was cleaned, it was found that the data types were inconsistent, so no conversion was needed. It looks very clean.



```
country       0.0000
child_mort    0.0000
exports       0.0000
health        0.0000
imports       0.0000
income        0.0000
inflation     0.0000
life_expec    0.0000
total_fer     0.0000
gdpp          0.0000
dtype: float64
```

```
country       object
child_mort    float64
exports       float64
health        float64
imports       float64
income        int64
inflation     float64
life_expec    float64
total_fer     float64
gdpp          int64
dtype: object
```

Figure 1. Cleaning results.

3.3. Data visualization

Next, the row visualization of the data set begins.

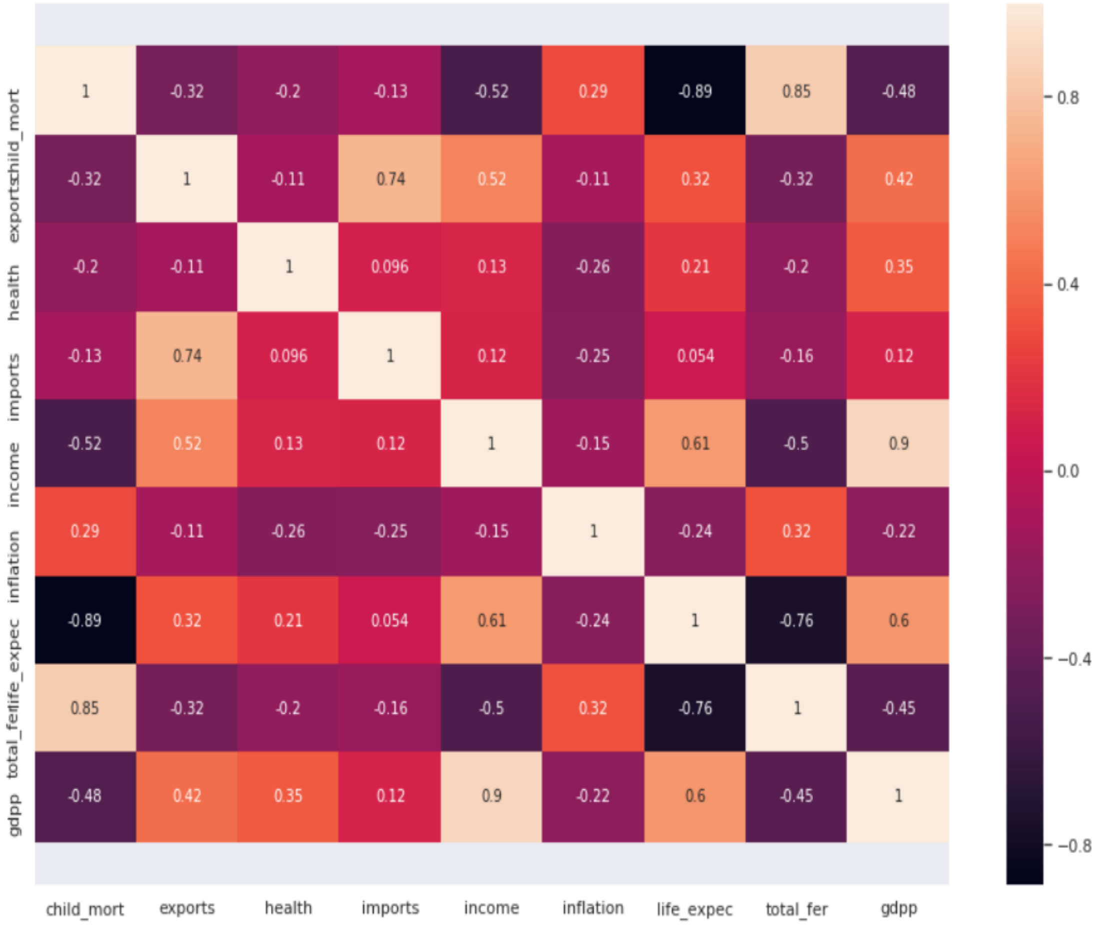


Figure 2. Heat map probe.

Correlation coefficient is to judge whether the data has useful value. The higher the correlation coefficient between each dimension in the data, the more reliable the data can be. According to Figure 2. The correlation coefficient between child mortality and life expectancy was -0.89. The correlation coefficient between child mortality and total fertility rate was 0.85. The correlation coefficient between import and export is 0.74. The correlation coefficient between life expectancy and total fertility rate was -0.76.

3.4. Data preparation

Because percentage values do not give a clear picture of the situation in the country. So we need to subtract imports, exports and health spending from their per capita GDP percentage values in real terms. Austria and Belarus, for example, have huge differences in GDP but the same share of exports. This situation does not accurately indicate which country is more developed than the other.

In the application analysis of this study, the rescale function Most software packages use SVD to calculate the principal component and assume that the data is scaled and centralized, so it is important to standardize or normalize. There are two common methods of scaling. Min-Max scaling and standardization (mean-O, sigma-1). In this case, we will use standardized scaling[9].

3.5. PCA application

In this study, PCA was used to remove redundancy in the data. The United Nations uses a similar heuristic to calculate the Human Development Index (HDI) to rank countries according to their development.

PAC is one of the most commonly used dimensionality reduction techniques in the industry. It improves model performance, visualizes complex data sets, and assists in a variety of other areas by converting large data sets to smaller ones with fewer variables.

Let's use PCA for dimensionality reduction, as it is obvious from the heat map that there is a correlation between the attributes.

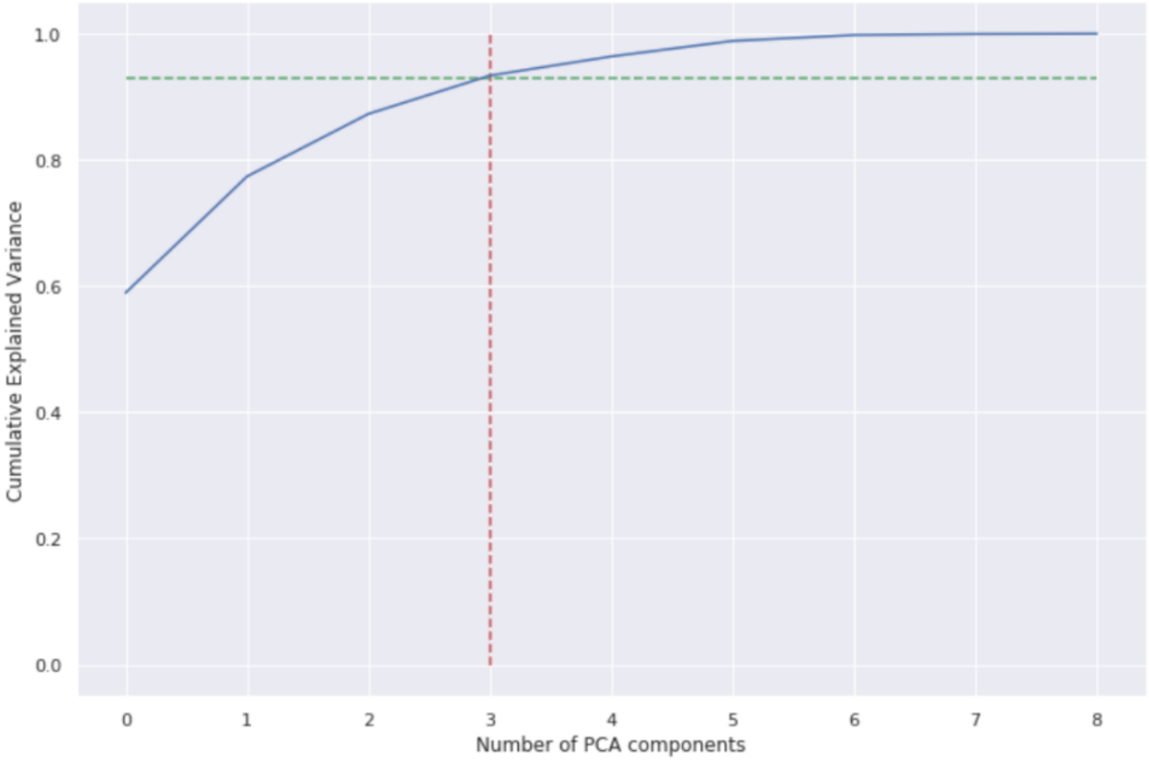


Figure 3. Visualize the number of components corresponding to the cumulative variance.

It is clear from Figure 3 above that more than 90% of the variance is explained by the first 3 principal components. Therefore, we will only use these components during the clustering process.

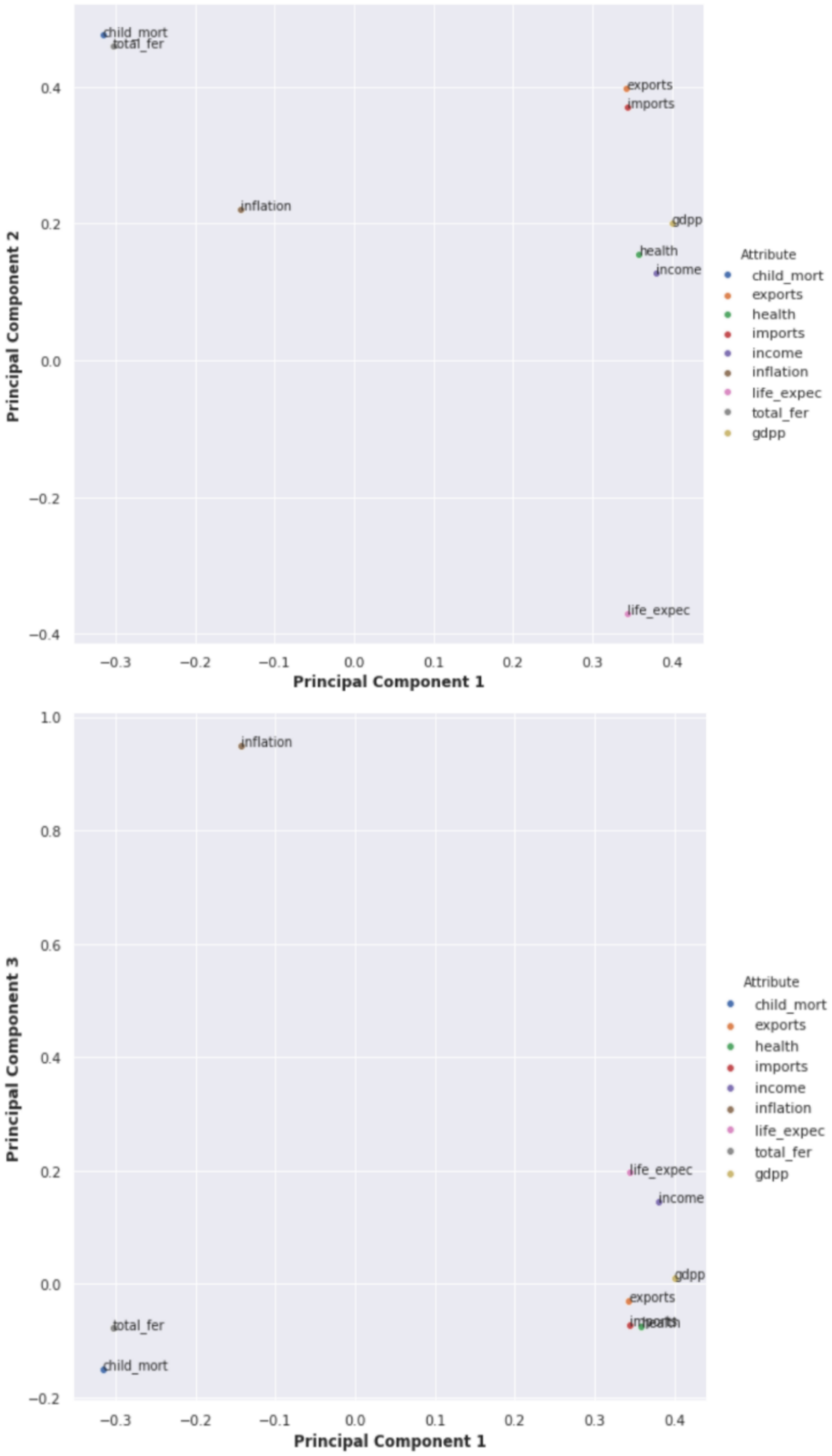


Figure 4. Data features data framework.

In Figure 4 as you can see from the two panels, PC1 does a good job of explaining life expectancy, income, GDP, and health status. Components PC1 and PC2 both explain imports and exports well. PC2 is a good explanation for child mortality and total fertility. Inflation can't be explained by PC1 and PC2. PC3 explains inflation very well. Since 90% of the variance is explained by three principal components, let's use only these three components to frame the data.

Table 2. Creating new data frame with Principal components.

	country	PC_1	PC_2	PC_3
0	Afghanistan	-2.6374	1.4690	-0.5414
1	Albania	-0.0223	-1.4319	-0.0207
2	Algeria	-0.4576	-0.6733	0.9619
3	Angola	-2.7245	2.1746	0.6067
4	Antigua and Barbuda	-0.6498	-1.0244	-0.2501

And statistical outliers can be found in table 2. We will deal with outliers because they string our data set. Finally, the data processing is shown in Figure 5.

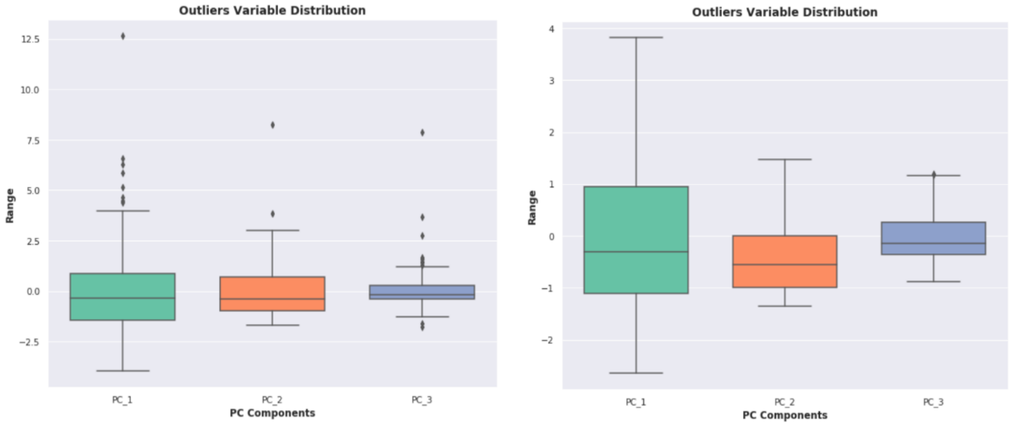


Figure 5. Plot after outlier removal.

3.6. Hopkins statistics test

The clustering tendency of the data set is measured according to Hopkins statistics as a test. The test score calculated in this study was 0.83. According to the decision rule, this is a good clustering Hopkins score.

3.7. Model building

The model works as follows: First, we randomly initialize k points as the mean value. We classify each item as the average closest to the mean and update the coordinates of the mean. We repeat a certain number of iterations, and finally, we get clustering. Elbow method is an intuitive method to determine the optimal value of k.

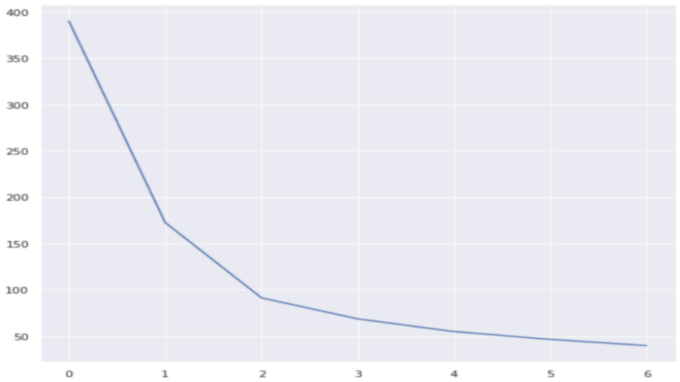


Figure 6. Elbow curve method.

Looking at the curve above, four or five clusters seem worth a try.
Before clustering, start assigning labels. Obtain Table 3

Table 3. Assigning labels.

	country	PC_1	PC_2	PC_3	Cluster_Id4
0	Afghanistan	-2.6374	1.4690	-0.5414	3
1	Algeria	-0.4576	-0.6733	0.9619	0
2	Antigua and Barbuda	-0.6498	-1.0244	-0.2501	2
3	Albania	-0.3327	-1.2745	0.1766	0
4	Australia	3.1804	-0.2508	-0.1169	1

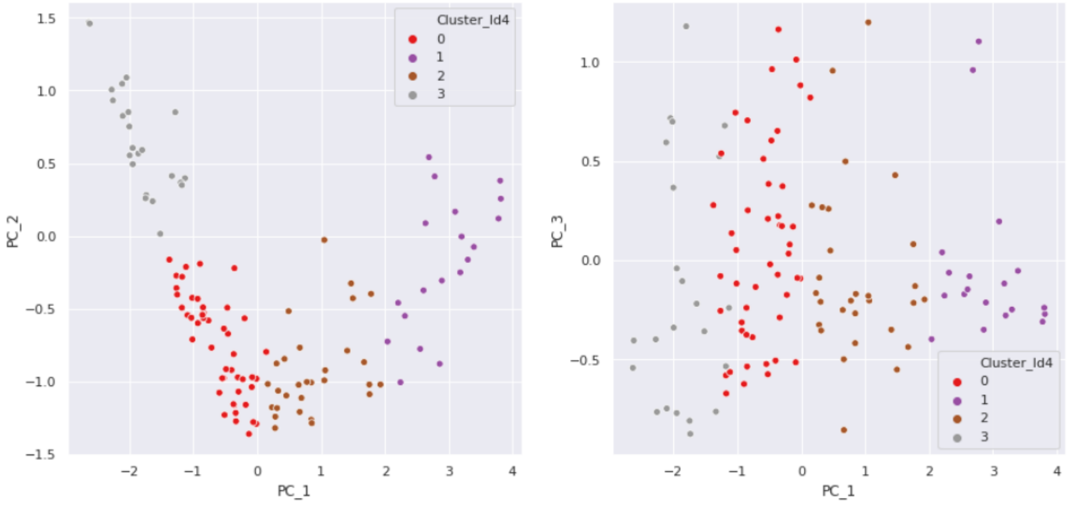


Figure 7. Visual analysis diagram of 4 clusters.

In Figure 7, there are quite a few countries in each cluster. However, it is also found that there are a lot of internal distances between cluster elements, which is not a good phenomenon.

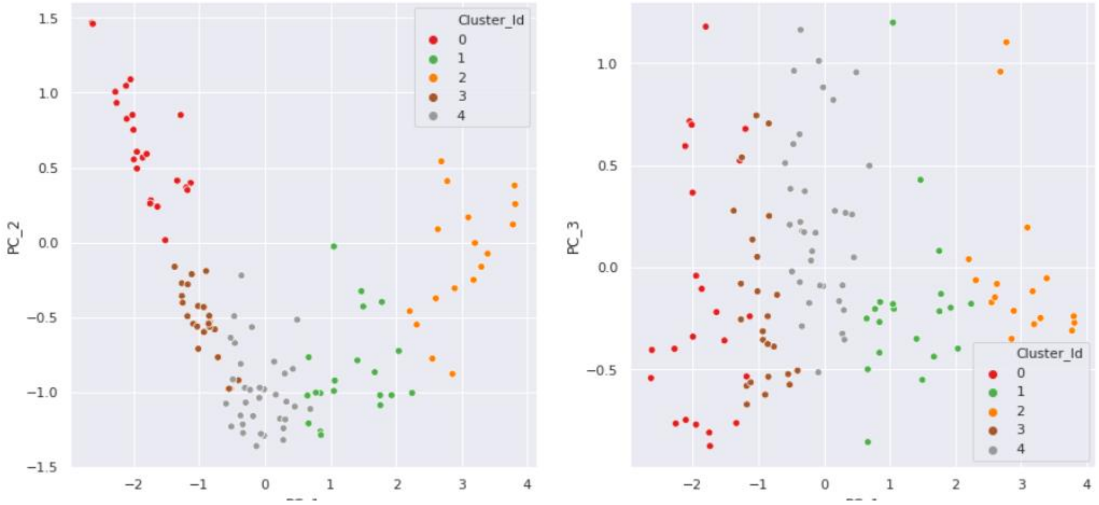


Figure 8. Visual analysis diagram of 5 clusters.

Here also, we have the same issue as with 4 clusters, but we have a new segment, so let's proceed with K means using 5 clusters.

We have visualized the data on the principal components and seen that some good clusters were formed, but others were not so good. Let's now visualize the data on the original attributes.

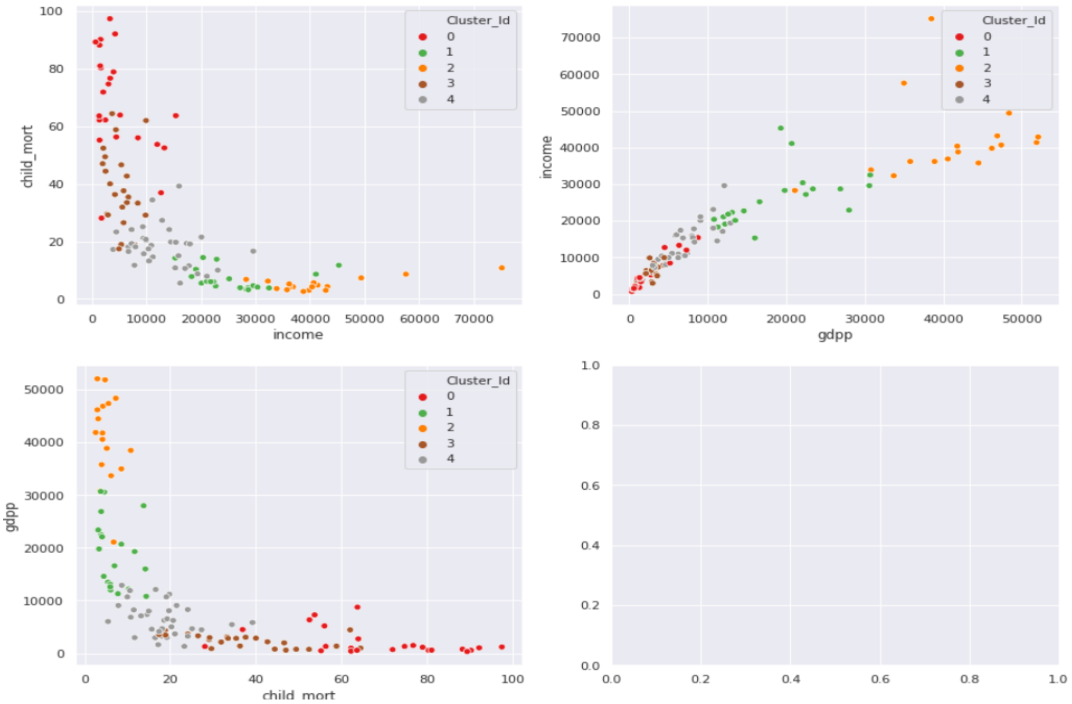


Figure 9. Scatter plots on raw attributes to visualize the distribution of data.

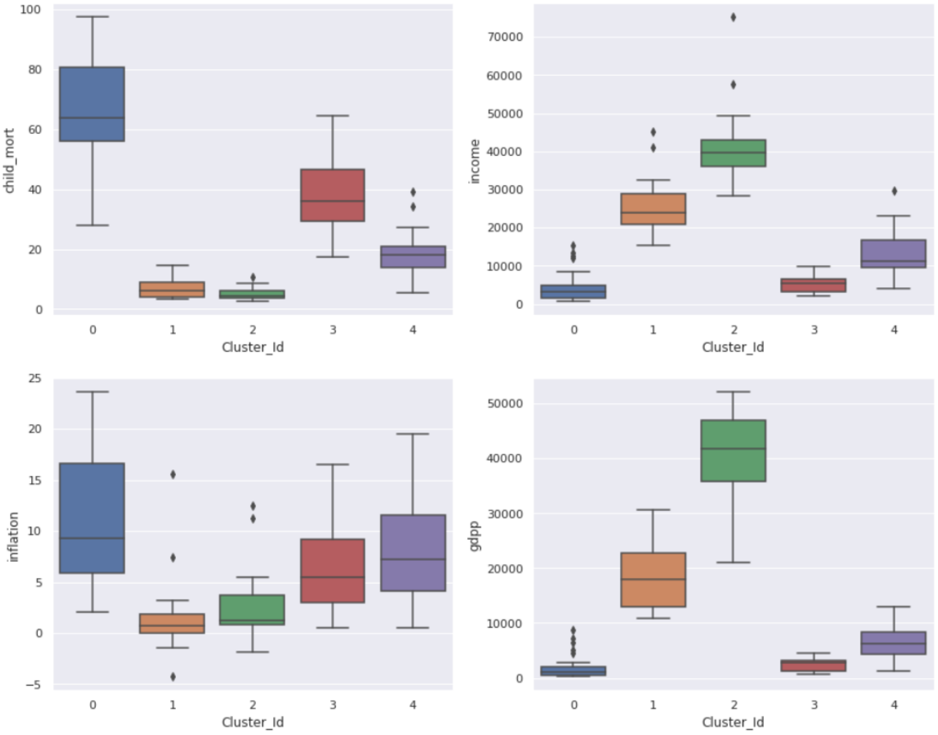


Figure 10. Box plot on original attributes to visualize the spread of the data.

According to figures 9 and 10, clusters 0 and 3 have the highest child mortality rates. These clusters need some assistance. Income and GDP are indicators of development. The higher the per capita income and GDP, the better the national development. Groups 0 and 3 appear to have the lowest per capita income and GDP. So, these countries need some help. Table 4 Box plot to visualize the mean value of few original attributes.

3.8. Final analysis

Using the PCA above to reduce the variables involved, countries are then clustered according to these major components. The country cluster is then constructed on the basis of identifying factors such as child mortality and income that are critical in determining the state of national development. Based on these categories, we have identified the following list of countries in urgent need of assistance. The list of countries/regions may change based on factors such as the selected survey dimension, the type of clustering, the clustering method used, etc. Of course, the model also exists in the process of outlier processing, which excludes some countries/regions. At the same time, the characteristics of social data will produce data errors. It may also exclude some countries/regions that may need help [10].

Table 4. Final countries list.

1	Afghanistan
2	Benin
3	Burkina Faso
4	Burundi
5	Central African Republic
6	Comoros
7	Congo, Dem. Rep.
8	Eritrea
9	Gambia
10	Guinea
11	Guinea-Bissau
12	Haiti
13	Liberia
14	Madagascar
15	Malawi
16	Mali
17	Mozambique
18	Niger
19	Rwanda
20	Sierra Leone
21	Tanzania
22	Togo
23	Uganda

4. Conclusion

To sum up, PCA was used in this study to reduce characteristic variables, and then K-mean clustering was performed for the countries after noise reduction according to the principal components of these less developed regions combined with Hopkins statistics. The multidimensional features ultimately target child mortality, income and so on, which are decisive in terms of the state of development of the country. On this basis, the country cluster is established. According to these categories, international NGOs can be organized to allocate resources more rationally and accelerate the rapid development of underdeveloped areas. However, this improved K-means clustering based on PCA also has some shortcomings. Although the combination of Hope skin statistics improves the clustering accuracy and breaks through the local optimization, it still cannot get rid of the dependence on the initial value setting.

In future research, the next step is to conduct multidimensional analysis and optimization of the anneal-based algorithm through research and experiments, aiming at the limitations of the unsupervised algorithm which depends on the original value J setting.

References

[1] Xu Z M." On the application of Machine Learning in Artificial Intelligence." Electronic World,2018,76-77.

[2] Han X W, Zhao T J." Machine Learning and Application of Imprecise Concepts." Journal of Harbin Institute of Technology,2006,122-125.

[3] Khan, Shejuti, Rahman, S.M. Monzurur, Tanim, M. Faysal, Ahmed, Fizar. (2013). Factors influencing K means algorithm. International Journal of Computational Systems Engineering, 217.doi:10.1504/ijcsyse.2013.057212.

[4] Zhang Y, Chen L" Study on Feature Extraction Method Based on Clustering and PCA Fusion." Computer Engineering and Applications,2010,152-154+193.

[5] MAO G J. The Concept, System Structure and Method of Data Mining [J]. Computer Engineering and Design, 2002,23 (8):13-17.

[6] Liu R Q." A Face recognition method with Occlusion based on PCA and HOG." Computer Knowledge and Technology,2019,177-179.

[7] Kutz J. N. Fu X X, Brunton S. L. Multi-resolution dynamic Modal decomposition. Journal of Applied Dynamic Systems, 15(2), pp. 713-735. 2016.

[8] He N, Ma M M." An improved K-means method for discovering Hot Topics in Weibo." Data Communication,2019,35-39.

[9] Brenton. Lee. Tu J H. Compressed sensing and dynamic modal decomposition. Journal of Computational Dynamics, 2015, 2(2).

[10] Boni, John, Li Y (Trans.), Li S (Trans.), Li H J, You F. (2018). Review of International NGO Development and Research. China Non-profit Review,22-53.